Kỷ yếu Hội nghị Khoa học công nghệ Quốc gia lần thứ XVI về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Đà Nẵng, ngày 28-29/9/2023
DOI: 10.15625/vap.2023.0089

XẤP XỈ ĐA THỨC TRONG TÁI TẠO ĐƯỜNG TẦN SỐ CƠ BẢN F0 CHO CÁC TỪ GHÉP TIẾNG VIỆT DỰA TRÊN MÔ HÌNH QTA VÀ THUẬT TOÁN TỐI ƯU HÀM MỤC TIÊU

Tạ Yên Thái[1], Vũ Thị Hải Hà[2], Đào Văn Tuyết[3], Ngô Hoàng Huy[4], Nguyễn Văn Hùng[5], Đoàn Văn Hòa[5], Trần Công Hùng[3]

[1]Đại học Kinh doanh và Công nghệ Hà Nội
[2]Phòng Ngữ âm - Từ vựng - Ngữ pháp, Viện Ngôn ngữ học,
Viện Hàn lâm Khoa học Xã hội Việt Nam
[3]Trường Đại học Quốc tế Sài gòn
[4]Trường Đại học CMC
[5]Viện Công nghệ thông tin, Viện Khoa học và Công nghệ quân sự

tayenthai@gmail.com, haihavu28@gmail.com, tuyetdv@gmail.com, nhhuy@cmc-u.edu.vn, nvhnt73@gmail.com, doanvanhoa@gmail.com, tranconghung@siu.edu.vn

TÓM TẮT: *Một trong những tham số cơ bản của tiếng nói sẽ được trích rút từ mô hình qTA (quantitative target approximation) để tái tạo ra các đường tần số cơ bản F0 cho các từ ghép tiếng Việt, tuy nhiên việc ước lượng tự động các tham số này là không dễ dàng do đường F0 có hình dạng phức tạp xuất hiện do hiện tượng biến thanh ở các câu trong ngữ cảnh có xuất hiện biến thanh. Sử dụng hướng tiếp cận xấp xỉ hàm liên tục bằng các đa thức, trong bài báo này, chúng tôi đề xuất một thuật toán tối ưu hàm mục tiêu để tự động ước lượng các tham số của mô hình qTA cho các âm là từ ghép tiếng Việt. Kết quả thực nghiệm cho thấy thuật toán được đề xuất có thể tái tạo đường F0 của các âm đôi tiếng Việt trong các ngữ cảnh câu nói khác nhau.*

Từ khóa: *Đường F0, Ước lượng F0, Mô hình Xu, qTA, Xấp xỉ đa thức,...*

I. GIỚI THIỆU

Theo phương pháp tổng hợp tiếng nói dựa trên phân tích đặc trưng, đã có nhiều nghiên cứu nhằm cải thiện khả năng tái tạo giọng nói. Mặc dù vậy, việc ước lượng và mô hình hóa đường cong F0 (tần số cơ bản) của âm thanh tiếng nói vẫn là những thách thức trong lĩnh vực này. Trong đó, một âm thanh có thể là tổ hợp của nhiều tần số, tần số chính được bao trùm trong âm được gọi là tần số cơ bản. Trong tiếng nói, tần số cơ bản là sự đáp ứng của sự rung động các dây thanh âm. Tần số cơ bản có giá trị phụ thuộc vào tần số lấy mẫu và khoảng cách giữa 2 đỉnh của các sóng âm tuần hoàn. Tổng hợp đường thanh điệu không chỉ làm cơ sở cho việc phát triển các hệ thống nhận diện thanh điệu mà còn có thể áp dụng rộng rãi trong lĩnh vực tổng hợp tiếng nói [17, 18].

Ngôn điệu (prosody) được hình thành từ tập hợp các yếu tố điệu tính gồm sự thay đổi cao độ, cường độ, tốc độ lời nói, trọng âm và cả âm sắc trong một đơn vị lời nói, đó là yếu tố quan trọng trong việc truyền đạt thái độ, nhấn mạnh sự thể hiện ngôn ngữ của mình tùy thuộc vào bối cảnh phục vụ giao tiếp hàng ngày. Trong quá trình phát triển trí tuệ nhân tạo, gần đây con người đang tìm cách giao tiếp với máy móc, hướng tới một giao tiếp tự nhiên gần giống con người hơn. Trong số các yếu tố tạo nên ngôn điệu (prosody), cao độ và tần số cơ bản là những đặc trưng quan trọng nhất.

Từ góc độ mô hình hóa, để tạo ra một mô hình có khả năng dự đoán đường nét F0 thì việc quan trọng là phải xác định các yếu tố dự đoán. Như đã đề xuất ở trên, nếu các yếu tố giao tiếp như trọng âm, ngữ điệu các loại câu và sự tương tác của chúng biểu hiện qua các đường cong F0 có hình dạng phức tạp, thì chúng sẽ là các yếu tố dự đoán. Một phương pháp thay thế cho mô hình hóa các yếu tố như vậy là mô phỏng đường cong F0 dựa trên các yếu tố dự đoán bao gồm các đặc điểm của mẫu F0 đã quan sát được, như cao độ, điểm chuyển đổi F0, và nhiều hơn nữa. Lý thuyết cho thấy mô hình hóa các yếu tố như vậy cung cấp một công cụ mạnh mẽ để kiểm tra giả thuyết về điệu nghệ và ngữ điệu. Quá trình như vậy được gọi là phân tích tổng hợp [1].

Giới công nghệ thông tin đã nỗ lực trong nhiều thập kỷ qua nhằm xây dựng một mô hình có khả năng mô phỏng các hiện tượng ngôn điệu khác nhau thông qua mô hình F0 [2, 3, 4, 5, 6]. Những phương pháp này có thể được chia thành hai loại chính: mô hình hóa F0 trực tiếp và mô phỏng cơ chế tái tạo F0.

Các mô hình thuộc loại đầu tiên như mô hình Fujisaki [3, 4, 5] chủ yếu được tạo ra dựa trên hình dạng của đường cong F0, với mục tiêu khớp các đường F0 từ dữ liệu đầu vào.

Loại thứ hai là mô hình mô phỏng cơ chế tái tạo F0 như qTA (Quantitative Target Approximation) [6], nhằm mô phỏng ngữ điệu, trọng âm và trọng tâm được dùng để nghiên cứu và phân tích giọng nói, làm rõ về cơ chế sản sinh giọng nói. qTA đã mô phỏng quá trình tạo ra giọng nói thông qua mô hình xấp xỉ mục tiêu định lượng để tạo ra đường cong F0 và tạo ra đường cong F0 cho giọng nói tiếng Anh và tiếng Trung Quốc.

Do đó để biểu diễn đường F0 của các từ ghép tiếng Việt, chúng tôi đã xác định theo hướng tiếp cận xấp xỉ đa thức để phân tích tham số của mô hình qTA.

Phần còn lại của bài báo được tổ chức như sau: Mục II Kiến thức liên quan, Mục III là thuật toán giải hàm với mục tiêu xác định tham số qTA dựa trên xấp xỉ đa thức, Mục IV là thực nghiệm và Mục V là kết luận.

II. NGHIÊN CỨU LIÊN QUAN VÀ PHƯƠNG PHÁP ĐỀ XUẤT

A. Mô hình qTA

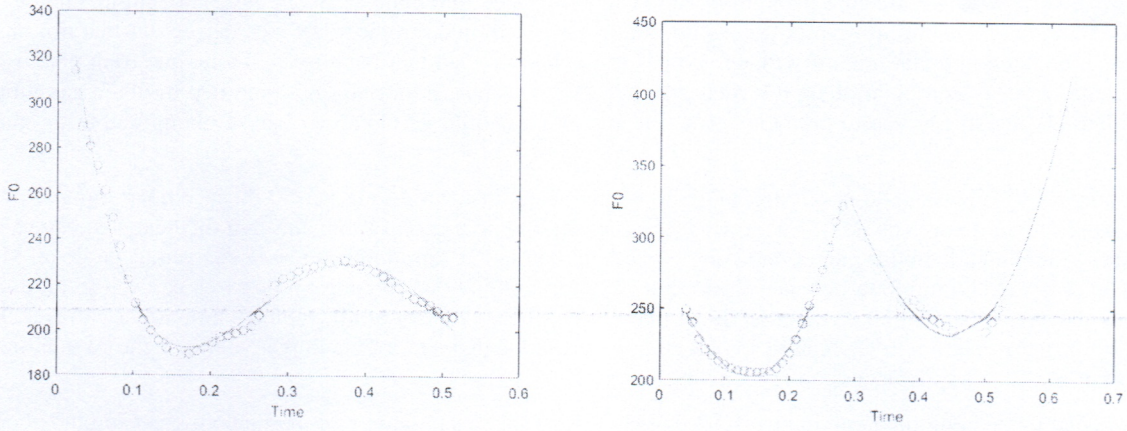
qTA (Xu [3-5], sử dụng một số tham số của giọng nói như đường nét tần số cơ bản F0, trường độ (duration) của từng đoạn, cao độ, thời điểm bắt đầu (onset timing) và thời điểm kết thúc (offset timing) âm tiết để tạo ra đường cong F0 cho giọng nói. Để đánh giá hiệu quả của mô hình qTA, các tham số mô hình được trích xuất từ giọng nói tự nhiên và so sánh với các giọng nói tự nhiên thông qua cả phân tích và kiểm tra của con người. Mô hình qTA có thể được sử dụng như một công cụ hiệu quả trong nghiên cứu về ngữ điệu và trọng âm, và cũng có thể tự động tạo ra ngữ điệu và trọng âm trong lời nói.

Với tiếng Anh, mô hình qTA sử dụng mô hình quan hệ thời gian và âm tiết để xác định thời điểm bắt đầu và kết thúc của các thành phần âm tiết. Đối với tiếng Mandarin, qTA sử dụng mô hình phụ thuộc âm tiết để xác định thời điểm bắt đầu và kết thúc của các thành phần âm tiết.

Mô hình đã được sử dụng rộng rãi cho tiếng Trung Quốc phổ thông-Mandarin, và trong [13] áp dụng cho tiếng Thái để mô hình hóa đường tần số cơ bản F0 của các thanh điệu trong ngữ cảnh. Từ một số nhỏ tham số của mô hình, chúng ta tạo sinh đường F0 mới có đường nét thích ứng cho các hiện tượng biến thanh của câu sẽ được phát âm. Các tác giả đã đề ra khái niệm "pitch target" trong tiếp cận mô hình hóa đường F0, trong thực tế ứng dụng thường "pitch target" là dạng tuyến tính, và về thực chất mô hình đã biểu diễn sai số xấp xỉ tuyến tính của các giá trị F0 bằng một xấp xỉ mới có dạng là hàm phân rã dạng hàm mũ có tốc độ suy giảm là một giá trị nằm trong khoảng (0,1) [6, 14].

B. Xấp xỉ đa thức

Định lý Weierstrass [7, 8, 16] về xấp xỉ hàm liên tục là một trong những định lý quan trọng nhất trong lý thuyết xấp xỉ hàm. Định lý này cho phép chúng ta xấp xỉ một hàm liên tục bất kỳ trên một đoạn đóng bằng cách sử dụng các đa thức. Giả sử chúng ta có một hàm f(x) liên tục trên đoạn đóng [a, b], trong đó a nhỏ hơn b. Định lý Weierstrass khẳng định rằng cho trước bất kỳ giá trị dương nào ϵ , chúng ta có thể tìm thấy một đa thức P(x) sao cho hiệu số tuyệt đối giữa f(x) và P(x) là nhỏ hơn ϵ trên toàn bộ đoạn [a, b], tức là $|f(x) - P(x)| < \epsilon$ với mọi x thuộc đoạn [a, b]. Định lý Weierstrass là một định lý tồn tại, cung cấp cho chúng ta một cách tiếp cận hữu ích để xấp xỉ và tính toán các hàm số phức tạp một cách hiệu quả về nguyên lý. Trong áp dụng với các hàm liên tục cụ thể, thì hoàn toàn có thể xây dựng dựng được đa thức xấp xỉ tốt của nó bằng các phương pháp được đề xuất bởi nhà Toán học Chebyshev. Ngoài phương pháp Chebyshev thì đa thức Legendre cũng được sử dụng để mô tả các đường để các điệu hóa đường F0 [9].



Hình 1. Khớp đường F0 bằng đa thức bậc 4: a) bằng một đa thức bậc 4 từ "cùng nhau"; b) bằng 2 đa thức bậc 2 cho từ "khoảng" và bậc 4 cho từ "đăng" của từ đôi "khoảng đăng"

C. Chỉ số đánh giá độ lệch 2 dãy số

Edit distance (khoảng cách chỉnh sửa) là một kỹ thuật được sử dụng để đo độ tương đồng giữa hai chuỗi. Nó đo lường số lần chỉnh sửa (thêm, xóa hoặc thay đổi ký tự) cần thiết để chuyển đổi một chuỗi thành chuỗi khác.

Công thức tính Edit distance dựa trên cách thức tìm đường đi ngắn nhất trong một ma trận. Ma trận này có kích thước (m+1) x (n+1), trong đó m và n là độ dài của hai chuỗi cần so sánh. Giá trị ở hàng i, cột j của ma trận tương ứng với khoảng cách chỉnh sửa giữa một phần tử của chuỗi nguồn với một phần tử của chuỗi đích.

D. Biểu diễn hàm mục tiêu cho thanh điệu

Đặt $G(t) = \dfrac{F_0(t) - (a*t+b)}{k^t}$, G(t) là hàm liên tục trên đoạn [T_1, T_2]. Theo định lý Weierstrass [7, 8], tồn tại đa thức P(t) sao cho $\max_{t\in[T_1,T_2]} |G(t) - P(t)| < \varepsilon$, nghĩa là

$$\forall t \in [T_1, T_2], \left|F_0(t) - at - b - P(t)k^t\right| < \varepsilon k^t < \varepsilon \text{ by } 0<k<1, \text{ tức là } \max_{t\in[T1,T2]} \left|F_0(t) - \{at + b + P(t)k^t\}\right| < \varepsilon.$$

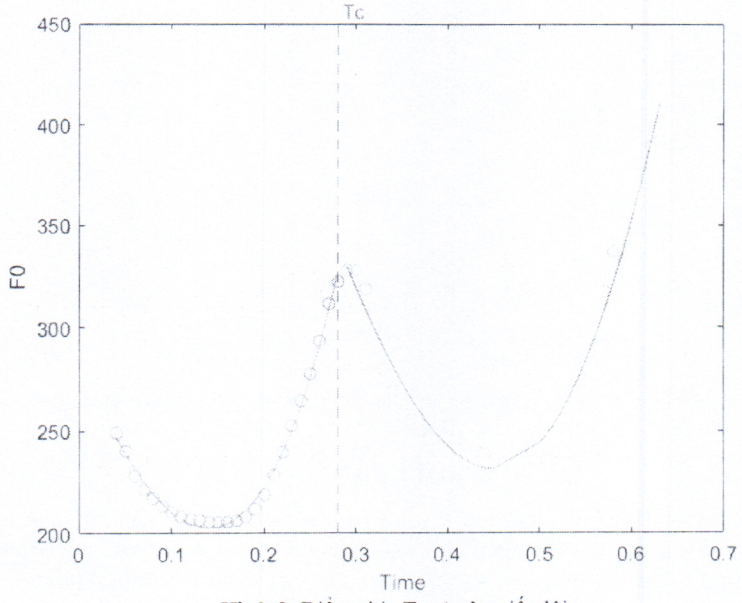
Đối với thanh điệu tn và đường F0 biểu diễn $t \mapsto f_0(t), 1 \le t \le T$ của tn, hàm mục tiêu PF được dự đoán các tham số của mô hình qTAV cho đường F0:

$$\begin{aligned} PF &\overset{def}{=} \beta*\Big(\sum_{T-last+1\le t\le T \wedge f_0(t)\neq NAN} \left\{f_0(t) - (a\,t+b)\right\}^2\Big) \\ &+\gamma*\Big(\sum_{1\le t\le T\wedge f_0(t)\neq NAN} \left\{f_0(t) - \left(a\ *t+b\ +P(t)*(k\ )^t\right)\right\}^2\Big) \to \min, k_{min} \le k \le k_{max} \end{aligned} \quad (1)$$

E. Thuật toán đề xuất

Với âm đôi thì điểm chia giữa các âm tiết sẽ được tính theo công thức dưới đây:

$T_c \overset{def}{=} \left[(T_1 + T_2)/2\right]$ T_1: Chỉ số frame cuối cùng của đường F0 nhanh thứ nhất âm, T_2 là chỉ số frame đầu tiên của âm thứ 2.



Hình 2. Điểm chia T_c của âm tiết đôi

Thuật toán trích rút tham số từ tập F0 cho từ đôi âm tiết.

Bước 1: Giải hàm mục tiêu PF là một xấp xỉ đa thức trên tập giá trị F0 từ frame thứ 1 đến frame $T_c - 1$. Sau khi giải hàm mục tiêu thì thu được bộ hệ số của đa thức xấp xỉ bậc không quá 4 (PC_1) và các hệ số a_1, b_1, k_1 để biểu diễn tham số mô hình tham số qTA.

Bước 2: Giải hàm mục tiêu PF là một xấp xỉ đa thức trên tập giá trị F0 từ frame thứ T_c đến frame cuối cùng của âm tiết thứ 2. Sau khi giải hàm mục tiêu thì thu được bộ hệ số của đa thức xấp xỉ bậc không quá 4 (PC_2) và các hệ số a_2, b_2, k_2 để biểu diễn tham số mô hình tham số qTA.

Bước 3: Tổng hợp 2 bước thì thu được 2 bộ hệ số hệ số PC_1, PC_2 và các hệ số $a_1, b_1, k_1, a_2, b_2, k_2$.

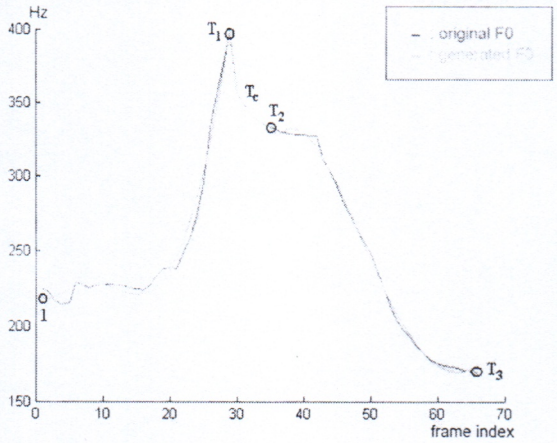
III. THỰC NGHIỆM

Bảng 1. Thứ tự các hệ số bậc đa thứ, a và k của qTAVs của ngữ điệu tiếng Việt với các hình dáng F0 khác nhau (Tổng quan, xu hướng/khoảng thời gian/hướng), 2' và 6' lần lượt là 2 và 6 ngữ điệu kết thúc bằng phụ âm đóng [10, 11]

Thanh điệu	Ví dụ	Hình F0	Bậc đa thức xấp xỉ	a	k
1	Ba	cao/ngang/dài/không đổi hướng	0	*	0
2	Má	cao/đi lên/dài/không đổi hướng	4	*	*
2'	Mắt	cao/đi lên/ngắn/không đổi hướng	0	0	0
3	Mã	cao/gãy/dài/đổi hướng đi xuống sau đi lên/tắc họng	4	*	*
4	Quả	thấp/cong/đổi hướng đi xuống sau đi lên bằng với cao độ ban đầu	4	*	*
5	Mà	thấp/đi xuống/không đổi hướng	0	*	0
6	Quạ	thấp/ đi xuống/ngắn/không đổi hướng	4	*	*
6'	Mặt	thấp/đi xuống/ngắn/không đổi hướng	0	0	0

*: Không ràng buộc.

Hơn nữa, để lựa chọn các tham số tối ưu, chúng ta có thể sử dụng thủ tục polyfit của Matlab [15].



Hình 3. Kết quả bằng thuật toán đề xuất với từ "dún dẩy" với bậc đa thức xấp xỉ bằng 4

IV. KẾT LUẬN

Trong bài báo này, chúng tôi đã đề xuất một thuật toán trích rút tham số của mô hình qTA tái tạo đường F0 dựa trên cách tiếp cận đa thức xấp xỉ cho âm đôi tiếng Việt. Phương pháp được áp dụng với mục tiêu tạo ra các đường F0 có hình dạng phức tạp áp dụng trong thể hiện F0 các từ đôi tiếng Việt. Kết quả thực nghiệm đã chứng minh rằng phương pháp đề xuất không chỉ hiệu quả trong việc tạo ra các đường F0 mong muốn mà còn có khả năng thích nghi với đa dạng các ngữ cảnh âm thanh.

Chúng tôi đã mô tả quá trình tiếp cận theo hướng qTA, trong đó chú trọng vào việc trích rút thông tin quan trọng từ tập F0 để tái tạo hình dạng của các từ đôi âm tiết. Bằng cách tận dụng tính tương quan giữa các tham số, thuật toán đã đạt được hiệu suất ấn tượng trong việc xấp xỉ và tái tạo các đường F0 phức tạp.

TÀI LIỆU THAM KHẢO

[1] Y. X. Santiham Prom-on, Fang Liu, "Functional Modeling of Tone Focus and Sentences Type in Mandarin Chinese," *ICPhS XVII*, pp. 1638-1641, 2011.

[2] G. Bailly and H. B. SFC, "A trainable prosodic model," *The Production of Speech*, edited by P.F. MacNeilage Springer-Verlag, New York, pp. 348-364, 2005.

[3] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing," *The Production of Speech*, edited by P.F. MacNeilage Springer-Verlag, New York, pp. 39-55, 1983.

[4] G. Kochanski and Shih, "Prosody modeling with soft templates," *Speech Communication*, 39, pp. 311-352, 2003.

[5] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Speech Communication*, 39, pp. 233-242, 1984.

[6] YiXu and Q. E. Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese," *Speech Communication,* vol. 33, pp. 319-337, 2001.

[7] M. H. Stone, "The Generalized Weierstrass Approximation Theorem," *Mathematics Magazine*, pp. 237-254, 1948.

[8] J. L. P. H. Gzyl, "The Weierstrass Approximation Theorem and Large Deviations," *The American Mathematical Monthly*, pp. 650-653, 1997.

[9] Remijsen B, *Tone, in The Oxford Research Encyclopedia of Linguistics Online.* Oxford University Press, 2016.

[10] Dung, Mixdorff, "Fujisaki Model based F0 contours in Vietnamese TTS," *Proc. Interspeech,* pp. 1429-1432, 2004.

[11] Hansjoerg Mixdorf, Nguyen Tien Dung, Luong Chi Mai, "Quantitative Analysis and Synthesis of Syllabic Tones in Vietnamese," *Proc. in EUROSPEECH,* pp. 177-180, 2003.

[12] Yi Xu, Santitham Prom-on, "Articulatory-functional modeling of speech prosody: A review," *Interspeech,* pp. 46-49, 2010.

[13] M. Leg´at, J. Matou˘sek, D. Tihelka, "On the Detection of Pitch Marks Using a Robust MultiPhase Algorithm," *Speech Communication,* vol. 53, pp. 552-566, 2011.

[14] Ching X. Xu, Yi Xu, and Li-Shi Luo, "A pitch target approximation model for f0 contours in mandarin," *Proc. of ICPhS,* 1999.

[15] Mathworks, Polynomial curve fitting, https://www.mathworks.com/help/matlab/math/polynomial-curve-fitting.html.

[16] R. Pachon, "Algorithms for polynomial and rational approximation," *Thesis,* 2010.

[17] Yoshihiro Yamazaki, Yuya Chiba, Takashi Nose, A. Ito , "Neural Spoken-Response Generation Using Prosodic and Linguistic Context for Conversational Systems," *Interspeech 30 August*, pp. 246-250, 2021.

[18] Ta Yen Thai, Hoang Ngo Huy, Dao Van Tuyet, S. Ablameyko, Nguyen Van Hung, Doan Van Hoa, "Tonal languages speech synthesis using an indirect pitch markers and the quantitative target approximation methods," *Journal of the Belarusian State University. Mathematics and Informatics*, pp. 105-121, 2019.

PHỤ LỤC

Mô hình Fujisaki

Do hướng tiếp cận tái tạo đường F0 dựa trên "fitting", mô hình Fujisaki khá phức tạp,
Mô hình sinh ra F0 theo công thức sau:

$$\log F_0(t) = \log F_b + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{aj} \{G_a(t - T_{1j}) - G_a(t - T_{2j})\} \quad (1)$$

$$G_p = \begin{cases} \alpha^2 t.\exp(-\alpha), & for \quad t \geq 0 \\ 0 \quad , & for \quad t < 0 \end{cases} \quad (2)$$

$$G_a = \begin{cases} \min[1-(1+\beta t).\exp(-\beta t)], & for \quad t \geq 0 \\ 0 \quad , & for \quad t < 0 \end{cases} \quad (3)$$

ở đây, các tham số của mô hình Fujisaki bao gồm: Ap, T0, a, Aa, T1, T2, β, Fb. Trong đó:

F_b - Giá trị cơ bản của tần số cơ bản,
I - số lệnh ngữ,
J - số lệnh trọng âm,
A_{pi} - cường độ của lệnh ngữ thứ i,
T_{0i} - thời điểm bắt đầu lệnh ngữ thứ i,
A_{aj} - biên độ của lệnh trọng âm thứ j,
T_{1j} - thời điểm bắt đầu thanh điệu ở lệnh trọng âm thứ j,
T_{2j} - thời điểm kết thúc thanh điệu ở lệnh trọng âm thứ j,
α - tần số góc tự nhiên của lệnh ngữ (phrase command) ,
β - tần số góc tự nhiên của lệnh ngữ (phrase command),
γ - mức giá trị trần tương ứng với các thành phần trọng âm.

POLYNOMIAL APPROXIMATION IN RECONSTRUCTING FUNDAMENTAL FREQUENCY (F0) CONTOURS FOR VIETNAMESE COMPOUND WORD TONES BASED ON THE QTA MODEL AND BASED ON THE QTA MODEL AND OPTIMAL OBJECTIVE FUNCTION ALGORITHM

Ta Yen Thai, Vu Thi Hai Ha, Dao Van Tuyet , Ngo Hoang Huy, Nguyen Van Hung, Doan Van Hoa, Tran Cong Hung

ABSTRACT: *One of the fundamental parameters of speech will be extracted from the qTA model (quantitative target approximation) to reproduce the fundamental frequency F0 contours for Vietnamese two-syllable tones. However, automatic estimation of these parameters is challenging due to the complex shape of the F0 contours caused by tonal sandhi in sentence contexts. Following the approach of approximating continuous functions with polynomials, in this paper, we propose an objective function optimization algorithm to automatically estimate the parameters of the qTA model for two-syllable tones of Vietnamese. Experimental results show that the proposed algorithm can reproduce the F0 contours of Vietnamese two-syllable tones in various sentence contexts.*

Keywords: *F0 contours, F0 estimation, Xu model, qTA, Polynomial approximation.*